What is claimed is:

1. A method of validating a byte sequence, the method comprising:

defining a plurality of states for the byte sequence;

designating one or more noise states from among the plurality of states;

generating a most probable state sequence for the byte sequence;

utilizing said state sequence to identify all noise in the byte sequence; and

localizing said noise in said noise states.

2. The method of claim 1 further comprising deleting said noise from the byte sequence.

3. The method of claim1 wherein an ASCII state is also designated as a noise state.

4. The method of claim 1 wherein said generating of a most probable state sequence comprises calculating $P(X_0 \ldots X_n \mid S_0 \ldots S_n)$, representing the conditional probabilities of said byte sequence given a state sequence.

5. The method of claim 4 wherein said calculating $P(X_0 \ldots X_n \mid S_0 \ldots S_n)$ comprises assigning a state label S_i to each i^{th} byte X_i of the byte sequence so as to maximize the equation:

$$P(X_0 \ldots X_N \mid S_0 \ldots S_N) = P_0(S_0)\left[\prod_{i=1}^{N} \overline{A}(S_i \mid S_{i-1})\right]\left[\prod_{i=0}^{N} \overline{B}(X_i \mid S_i)\right]$$

wherein $P_0(S_0)$ is the initial distribution of states; $\overline{A}(S_i \mid S_{i-1})$ is a "state-to-state" transmission matrix; and $\overline{B}(X_i \mid S_i)$ is a "byte-from-state" matrix of the probabilities of generating a byte value X_i given a state S_i.

6. The method of claim 5, wherein:

$$\overline{A}(S_i \mid S_{i-1}) = \begin{bmatrix} p(S_{i-1}^1 \rightarrow S_i^1) & \cdots & p(S_{i-1}^1 \rightarrow S_i^\sigma) \\ \vdots & \ddots & \vdots \\ p(S_{i-1}^\sigma \rightarrow S_i^1) & \cdots & p(S_{i-1}^\sigma \rightarrow S_i^\sigma) \end{bmatrix}$$

where each $p(S_{i-1} \rightarrow S_i)$ is the probability that a particular S_i state immediately follows an S_{i-1} state in a valid byte sequence having σ states.

7. The method of claim 8, wherein:

$$\overline{A}(S_i \mid S_{i-1}) = \begin{bmatrix} p(A \rightarrow A) & p(A \rightarrow GB1) & p(A \rightarrow GB2) \\ p(GB1 \rightarrow A) & p(GB1 \rightarrow GB1) & p(GB1 \rightarrow GB2) \\ p(GB2 \rightarrow A) & p(GB2 \rightarrow GB1 & p(GB2 \rightarrow GB2) \end{bmatrix}$$

where each $p(S_{i-1} \rightarrow S_i)$ is the probability that a particular S_i state immediately follows an S_{i-1} state in a valid byte sequence having three states.

8. The method of claim 7, wherein:

$$\overline{A}(S_i \mid S_{i-1}) = \begin{bmatrix} 0.995157 & 0.004843 & 0 \\ 0 & 0 & 1 \\ 0.037944 & 0.962056 & 0 \end{bmatrix}$$

and said valid byte sequence is valid text in the GB 2312-80 character set.

9. The method of claim 5, wherein:

$$\overline{B}(X_i \mid S_i) = \begin{bmatrix} h_1(X_i = 1) & \cdots & h_s(X_i = 1) & \cdots & h_\sigma(X_i = 1) \\ \vdots & \ddots & \vdots & & \vdots \\ h_1(X_i = x_1) & & h_s(X_i = x_1) & & h_\sigma(X_i = x_1) \\ h_1(X_i = x_1 + 1) & & \varepsilon_1(X_i = x_1 + 1) & & h_\sigma(X_i = x_1 + 1) \\ \vdots & & \vdots & & \vdots \\ h_1(X_i = x_r) & & \varepsilon_1(X_i = x_r) & & h_\sigma(X_i = x_r) \\ h_1(X_i = x_r + 1) & & \varepsilon_r(X_i = x_r + 1) & \ddots & h_\sigma(X_i = x_r) \\ \vdots & & \vdots & \ddots & \vdots \\ h_1(X_i = x_r = 255) & & \varepsilon_r(X_i = x_r = 255) & & h_\sigma(X_i = 255) \end{bmatrix}$$

where $h_S(X_i)$ are histogram functions of the σ states and $\varepsilon_j(X_i)$ are probabilities of associating noise with the noise state for bytes within $r+1$ ranges of byte values X_i.

10. The method of claim 9, wherein:

$$\overline{B}(X_i \mid S_i) = \begin{bmatrix} h_A(X_i = 1) & 0 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ h_A(X_i = 127) & 0 & 0 \\ \varepsilon_1(X_i = 128) & 0 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \varepsilon_1(X_i = 160) & 0 & 0 \\ \varepsilon_2(X_i = 161) & h_1(X_i = 161) & h_2(X_i = 161) \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ \varepsilon_2(X_i = 254) & h_1(X_i = 254) & h_2(X_i = 254) \\ \varepsilon_3(X_i = 255) & 0 & 0 \end{bmatrix}$$

where $h_s(X_i)$ are histogram functions of the states, and $\varepsilon_j(X_i)$ are probabilities of associating noise with the ASCII state within a plurality of X_i ranges for a three-state byte sequence.

11. The method of claim 10 further comprising:

providing a valid three-state byte sequence having an ASCII state and comprising valid ASCII and two-byte characters;

computing an ASCII histogram $h_A(X_i)$ by a method comprising:

sampling valid ASCII text so as to measure the frequency of occurrence of each byte value;

computing an unnormalized ASCII histogram of said sampling over the ASCII range of X_i; and

normalizing said unnormalized ASCII histogram such that the entire column of the matrix containing said ASCII histogram sums to 1;

computing a first-byte histogram $h_1(X_i)$ by sampling valid two-byte text and computing the unnormalized first-byte histogram over the odd bytes, and normalizing said first-byte histogram such that the entire column of the matrix containing said first-byte histogram sums to 1; and

computing a second-byte histogram $h_2(X_i)$ by sampling valid two-byte text and computing the unnormalized second-byte histogram over the odd bytes, and normalizing said second-byte histogram such that the entire column of the matrix containing said second-byte histogram sums to 1.

12. A program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for validating a byte sequence, said method comprising:

defining a plurality of states for the byte sequence;

designating one or more noise states from among the plurality of states;

generating a most probable state sequence for the byte sequence;

utilizing said state sequence to identify all noise in the byte sequence; and

localizing said noise in said noise states.

13. The device of claim 12 wherein said localizing of said noise in said noise states comprises:

examining each byte in said byte sequence that does not correspond to a noise state;

determining if the byte is valid; and

if the byte is not valid, then redesignating the state of said byte to a noise state.

14. The device of claim 13 further comprising:

a lookup table of valid bytes; and

wherein said determination if a byte is valid is accomplished by accessing said lookup table.

15. A method of validating a byte sequence, the method comprising:

defining a plurality of states for the byte sequence, including at least one ASCII state;

designating at least one ASCII state as the noise state;

generating a most probable state sequence for the byte sequence by a method comprising:

calculating $P(X_0 \ldots X_n \mid S_0 \ldots S_n)$, representing the conditional probabilities of said byte sequence given a state sequence;

wherein said calculating $P(X_0 \ldots X_n \mid S_0 \ldots S_n)$ comprises assigning a state label S_i to each i^{th} byte X_i of the byte sequence so as to maximize the equation:

$$P(X_0 \ . \ . \ . \ X_N \mid S_0 \ . \ . \ . \ S_N) = P_0(S_0) \left[\prod_{i=1}^{N} \overline{A}(S_i \mid S_{i-1})\right]\left[\prod_{i=0}^{N} \overline{B}(X_i \mid S_i)\right]$$

wherein $P_0(S_0)$ is the initial distribution of states; $\overline{A}(S_i \mid S_{i-1})$ is a "state-to-state" transmission matrix; and $\overline{B}(X_i \mid S_i)$ is a "byte-from-state" matrix of the probabilities of generating a byte value X_i given a state S_I;

utilizing said state sequence to identify all noise in the byte sequence;

localizing said noise in said noise states; and

deleting said noise from the byte sequence.